

# Error Bounds for Compressed Sensing Algorithms With Group Sparsity: A Unified Approach

M. Eren Ahsen and M. Vidyasagar \*

December 31, 2015

## Abstract

In compressed sensing, in order to recover a sparse or nearly sparse vector from possibly noisy measurements, the most popular approach is  $\ell_1$ -norm minimization. Upper bounds for the  $\ell_2$ - norm of the error between the true and estimated vectors are given in [1] and reviewed in [2], while bounds for the  $\ell_1$ -norm are given in [3]. When the unknown vector is not conventionally sparse but is “group sparse” instead, a variety of alternatives to the  $\ell_1$ -norm have been proposed in the literature, including the group LASSO, sparse group LASSO, and group LASSO with tree structured overlapping groups. However, no error bounds are available for any of these modified objective functions. In the present paper, a unified approach is presented for deriving upper bounds on the error between the true vector and its approximation, based on the notion of decomposable and  $\gamma$ -decomposable norms. The bounds presented cover all of the norms mentioned above, and also provide a guideline for choosing norms in future to accommodate alternate forms of sparsity.

## 1 Introduction

The field of “compressed sensing” has become very popular in recent years, with an explosion in the number of papers. In the interests of brevity, we

---

\*MEA is with IBM Research, Thomas J. Watson Research Center, 1101 Route 134 Kitchawan Rd, Yorktown Heights, NY 10598. MV is with the Systems Engineering Department, University of Texas at Dallas, Richardson, TX 75080. Emails: mahsen@us.ibm.com, m.vidyasagar@utdallas.edu. This research was Supported by the National Science Foundation under Award #1306630, and the Cecil H. & Ida Green Endowment at UT Dallas

refer the reader to two recent papers [2, 4], each of which contains an extensive bibliography. Stated briefly, the core problem in compressed sensing is to approximate a high-dimensional sparse (or nearly sparse) vector  $x$  from a small number of linear measurements of  $x$ . Though this problem has a very long history (see the discussion in [2] for example), perhaps it is fair to say that much of the recent excitement has arisen from [5], in which it is shown that if  $x$  has no more than  $k$  nonzero components, then by choosing the matrix  $A$  to satisfy a condition known as the restricted isometry property (RIP), it is possible to recover  $x$  exactly by minimizing  $\|z\|_1$  subject to the constraint that  $Az = y = Ax$ . In other words, under suitable conditions, among all the preimages of  $y = Ax$  under  $A$ , the preimage that has minimum  $\ell_1$ -norm is the sparse signal  $x$  itself. The same point is also made in [6]. Subsequently the RIP was replaced by the null space property [7], which is actually necessary and sufficient for the above statement to be true; see [3, Chapter 4] for precise statements. In case  $y = Ax + \eta$  where  $\eta$  is a measurement error and  $x$  is either sparse or nearly sparse, one can attempt to recover  $x$  by setting

$$\hat{x} := \operatorname{argmin}_{z \in \mathbb{R}^n} \|z\|_1 \text{ s.t. } \|y - Az\|_2 \leq \epsilon. \quad (1)$$

This algorithm is very closely related to the LASSO algorithm introduced in [8]. Specifically, the only difference between LASSO as in [8] and the problem stated above is that the roles of the objective function and the constraint are reversed. It is shown (see [1, Theorem 1.2]) that, under suitable conditions, the residual error  $\|\hat{x} - x\|_2$  satisfies an estimate of the form

$$\|\hat{x} - x\|_2 \leq \frac{C_0}{\sqrt{k}} \sigma_k(x, \|\cdot\|_1) + C_2 \epsilon, \quad (2)$$

where  $\sigma_k(x, \|\cdot\|_1)$  is the “sparsity index” of  $x$  (defined below), and  $C_0, C_2$  are constants that depend only on the matrix  $A$  but not  $x$  or  $\eta$ . The above bound includes exact signal recovery with noiseless measurements as a special case, and is referred to in [1] as “noisy recovery.” Along similar lines, it is shown in [3] that

$$\|\hat{x} - x\|_1 \leq C_0 \sigma_k(x, \|\cdot\|_1) + C_2 \sqrt{k} \epsilon, \quad (3)$$

where  $C_0$  and  $C_2$  are the same as in (2). See the equation just above [3, Equation (4.16)].

In the world of optimization, the LASSO algorithm has been generalized in several directions, by modifying the  $\ell_1$ -norm penalty of LASSO to some other norm that is supposed to induce a prespecified sparsity structure on

the solution. Among the most popular sparsity-inducing penalty norms are the group LASSO [9, 10], referred to hereafter as GL, and the sparse group LASSO [11, 12], referred to hereafter as SGL. Now there are versions of these algorithms that permit the groups to have “tree-structured” overlap [13, 14].

It is therefore natural to ask whether inequalities analogous to (2) and (3) hold when the  $\ell_1$ -norm in (1) is replaced by other sparsity-inducing norms such as those mentioned in the previous paragraph. To the best of the authors’ knowledge, no such error bounds are available in the literature for anything other than  $\ell_1$ -norm minimization. In principle, it is possible to mimic the arguments in [1] to derive error bounds for each of these algorithms. However, it would be highly desirable to have a unified theory of what properties a norm needs to satisfy, in order that inequalities of the form (2) hold. That is the focus of the present paper. We present a very general result to the effect that *any* compressed sensing algorithm satisfies error bounds of the form (2) and (3) provided three conditions are satisfied:

1. A “compressibility condition” holds, which in the case of  $\ell_1$ -norm minimization is that the restricted isometry property (RIP) holds with a sufficiently small constant.
2. The approximation norm used to compute the sparsity index of the unknown vector  $x$  is “decomposable” as defined subsequently.
3. The penalty norm used to induce the sparsity of the solution, that is, the norm that is minimized, is “ $\gamma$ -decomposable” as defined subsequently.

It will follow as a consequence of this general result that GL, and SGL (without or with tree-structured overlapping groups) all satisfy error bounds of the form (2). In addition to the generality of the results established, the method of proof is more direct than that in [1, 2]. In the case of conventional sparsity and  $\ell_1$ -norm minimization, the results presented here contain those in [1, 2] as special cases, and also include a bound on  $\|\hat{x} - x\|_1$ , in addition the bound on  $\|\hat{x} - x\|_2$ .

## 2 Preliminaries

If  $x \in \mathbb{R}^n$ , and  $\Lambda$  is a subset of  $\mathcal{N} = \{1, \dots, n\}$ , the symbol  $x_\Lambda \in \mathbb{R}^n$  denotes the vector such that  $(x_\Lambda)_i = x_i$  if  $i \in \Lambda$ , and  $(x_\Lambda)_i = 0$  if  $i \notin \Lambda$ . In other

words,  $x_\Lambda$  is obtained from  $x$  by replacing  $x_i$  by zero whenever  $i \notin \Lambda$ . Also, as is customary, for a vector  $u \in \mathbb{R}^n$ , its support set is defined by

$$\text{supp}(u) := \{i : u_i \neq 0\}.$$

Let  $k$  be some integer that is fixed throughout the paper. Next we introduce the notion of a group  $k$ -sparse set. Some care is required in doing so, as the discussion following the definition shows.

**Definition 1.** Let  $\mathcal{G} = \{G_1, \dots, G_g\}$  be a partition of  $\mathcal{N} = \{1, \dots, n\}$ , such that  $|G_i| \leq k$  for all  $i$ . If  $S \subseteq \{1, \dots, g\}$ , define  $G_S := \cup_{i \in S} G_i$ . A subset  $\Lambda \subseteq \mathcal{N}$  is said to be **group  $k$ -sparse** if there exists a subset  $S \subseteq \{1, \dots, g\}$  such that  $\Lambda = G_S$ , and in addition,  $|\Lambda| \leq k$ . The collection of all group  $k$ -sparse subsets of  $\mathcal{N}$  is denoted by  $\text{GkS}$ . A vector  $u \in \mathbb{R}^n$  is said to be **group  $k$ -sparse** if its support set  $\text{supp}(u)$  is contained in a group  $k$ -sparse set.

At this point the reader might ask why a set  $\Lambda$  cannot be defined to be group  $k$ -sparse if it is a *subset* of some  $G_S$ , as opposed to being *exactly equal* to some  $G_S$ . The reason is that, if every subset of  $G_S$  is also called “group  $k$ -sparse,” then in effect *all sets of cardinality  $k$  or less* can be called group  $k$ -sparse, thus defeating the very purpose of the definition. To see this, let  $\Lambda = \{x_{i_1}, \dots, x_{i_l}\}$ , where  $l \leq k$ , so that  $|\Lambda| = l \leq k$ . Then, since the sets  $G_1, \dots, G_g$  partition the index set  $\mathcal{N}$ , for each  $j$  there exists a set  $G_j$  such that  $x_{i_j} \in G_j$ . Let  $S \subseteq \{1, \dots, g\}$  denote the set consisting of all these indices  $j$ . Then  $\Lambda \subseteq G_S$ . So with this modified definition, there would be no difference between group  $k$ -sparsity and conventional sparsity. This is the reason for adopting the above definition. On the other hand, it is easy to see that if  $g = n$  and each set  $G_i$  equals the singleton set  $\{i\}$ , then group  $k$ -sparsity reduces to conventional  $k$ -sparsity. Note also that a vector is defined to be group  $k$ -sparse if its support *is contained in*, though not necessarily equal to, a group  $k$ -sparse subset of  $\mathcal{N}$ .

Suppose  $\|\cdot\| : \mathbb{R}^n \rightarrow \mathbb{R}_+$  is some norm. We introduce a couple of notions of decomposability that build upon an earlier definition from [4].

**Definition 2.** The norm  $\|\cdot\|$  is said to be **decomposable** with respect to the partition  $\mathcal{G}$  if, whenever  $u, v \in \mathbb{R}^n$  with  $\text{supp}(u) \subseteq G_{S_u}$ ,  $\text{supp}(v) \subseteq G_{S_v}$ , and  $S_u, S_v$  are disjoint subsets of  $\{1, \dots, g\}$ , it is true that

$$\|u + v\| = \|u\| + \|v\|. \quad (4)$$

As pointed out in [4], because  $\|\cdot\|$  is a norm, the relationship (4) *always* holds with  $\leq$  replacing the equality. Therefore the essence of decomposability is that the bound is tight when the two summands are vectors with their support sets contained in disjoint sets of the form  $G_{S_u}, G_{S_v}$ . Note that it is not required for (4) to hold for every pair of vectors with disjoint supports, only vectors whose support sets are contained in disjoint unions of group  $k$ -sparse subsets of  $\mathcal{N}$ . For instance, if  $\Lambda$  is a group  $k$ -sparse set, and  $u, v$  have disjoint support sets  $\text{supp}(u), \text{supp}(v)$  that are both subsets of  $\Lambda$ , then there is no requirement that (4) hold. It is easy to see that the  $\ell_1$ -norm is decomposable, and it is shown below that the group LASSO and the sparse group LASSO norm are also decomposable. To generalize our analysis, we introduce a more general definition of decomposability.

**Definition 3.** *The norm  $\|\cdot\|$  is  $\gamma$ -decomposable with respect to the partition  $\mathcal{G}$  if there exists  $\gamma \in (0, 1]$  such that, whenever  $u, v \in \mathbb{R}^n$  with  $\text{supp}(u) \subseteq G_{S_u}$ ,  $\text{supp}(v) \subseteq G_{S_v}$ , and  $S_u, S_v$  are disjoint subsets of  $\{1, \dots, g\}$ , it is true that*

$$\|u + v\| \geq \|u\| + \gamma\|v\|. \quad (5)$$

Note that if the norm  $\|\cdot\|$  is  $\gamma$ -decomposable with  $\gamma = 1$ , then (5) and the triangle inequality imply that

$$\|u + v\| \geq \|u\| + \|v\| \implies \|u + v\| = \|u\| + \|v\|.$$

Therefore decomposability is the same as  $\gamma$ -decomposability with  $\gamma = 1$ .

Clearly, if  $\|\cdot\|$  is a decomposable norm, then (4) can be applied recursively to show that if  $\Lambda_0, \Lambda_1, \dots, \Lambda_s$  are pairwise disjoint group  $k$ -sparse sets, and  $\text{supp}(u_i) \subseteq \Lambda_i$ , then

$$\left\| \sum_{i=0}^s u_i \right\| = \sum_{i=0}^s \|u_i\|. \quad (6)$$

However, such an equality does not hold for  $\gamma$ -decomposable functions unless  $\gamma = 1$ , which makes the norm decomposable. On the other hand, by repeated application of (5) and noting that  $\gamma \leq 1$ , we arrive at the following relationship: if  $\Lambda_0, \Lambda_1, \dots, \Lambda_s$  are pairwise disjoint group  $k$ -sparse sets, and  $\text{supp}(u_i) \subseteq \Lambda_i$ , then

$$\left\| \sum_{i=0}^s u_i \right\| \geq \|u_{\Lambda_0}\| + \gamma \left\| \sum_{i=1}^s u_i \right\|. \quad (7)$$

Equation (6) is somewhat more general than the definition of decomposability given in [4], in that we permit the partitioning of the index set  $\mathcal{N}$  into more than two subsets. However, this is a rather minor generalization.<sup>1</sup>

It is now shown that the notions of decomposability and  $\gamma$ -decomposability are general enough to encompass several algorithms such as group LASSO, sparse group LASSO, either without overlapping groups or with groups that overlap but have a tree structure.

**Lemma 1.** *Let  $\mathcal{G} = \{G_1, \dots, G_g\}$  be a partition of the index set  $\mathcal{N} = \{1, \dots, n\}$ . Let  $\|\cdot\|_i : \mathbb{R}^{|G_i|} \rightarrow \mathbb{R}_+$  be any norm, and define the corresponding norm on  $\mathbb{R}^n$  by*

$$\|x\|_A = \sum_{i=1}^g \|x_{G_i}\|_i. \quad (8)$$

*Then the above norm is decomposable.*

The proof is omitted as it is obvious. The key point to note is that the exact nature of the individual norms  $\|\cdot\|_i$  is entirely irrelevant.

By defining the individual norms  $\|\cdot\|_i$  appropriately, it is possible to recover the group LASSO [9, 10], the sparse group LASSO [11, 12], and the overlapping sparse group LASSO with tree-structured norms [13, 14].

**Corollary 1.** *The group LASSO norm defined by*

$$\|z\|_{\text{GL}} := \sum_{i=1}^g \|z_{G_i}\|_2. \quad (9)$$

*is decomposable.*

**Proof:** This corresponds to the choice  $\|\cdot\|_i = \|\cdot\|_2$ . Note that some authors use  $\|z_{G_i}\|_2/\sqrt{|G_i|}$  instead of just  $\|z_{G_i}\|_2$ . This variant is also decomposable, as is easy to see.  $\square$

**Corollary 2.** *The sparse group LASSO norm defined by*

$$\|z\|_{\text{SGL}, \mu} := \sum_{i=1}^g [(1 - \mu)\|z_{G_i}\|_1 + \mu\|z_{G_i}\|_2]. \quad (10)$$

*is decomposable.*

---

<sup>1</sup>There is a little bit of flexibility in [4] in that one can take two orthogonal subspaces that are not exactly orthogonal complements of each other; but we will not belabor this point.

**Proof:** This corresponds to the choice

$$\|z_{G_i}\|_i = (1 - \mu)\|z_{G_i}\|_1 + \mu\|z_{G_i}\|_2.$$

Therefore the norm is decomposable.  $\square$

Next let us turn our attention to the case of “overlapping” groups with tree structure, as defined in [13, 14].

**Corollary 3.** *Suppose there are sets  $\mathcal{N}_1, \dots, \mathcal{N}_l$ , each of which is a subset of  $\mathcal{N}$ , that satisfy the condition*

$$\mathcal{N}_i \cap \mathcal{N}_j \neq \emptyset \implies (\mathcal{N}_i \subseteq \mathcal{N}_j \text{ or } \mathcal{N}_j \subseteq \mathcal{N}_i). \quad (11)$$

Define

$$\|z\|_A = \sum_{i=1}^l \|z_{\mathcal{N}_i}\|_i,$$

where  $\|\cdot\|_i : \mathbb{R}^{|\mathcal{N}_i|} \rightarrow \mathbb{R}_+$  is arbitrary. Then this norm is decomposable.

**Proof:** Though it is possible for some of these sets  $\mathcal{N}_i$  to overlap, the condition (11) implies that the collection of sets  $\mathcal{N}_1, \dots, \mathcal{N}_l$  can be renumbered with double indices as  $\mathcal{S}_{ij}$ , and arranged in chains of the form

$$\mathcal{S}_{11} \subseteq \dots \subseteq \mathcal{S}_{1n_1}, \dots, \mathcal{S}_{s1} \subseteq \dots \subseteq \mathcal{S}_{sn_s},$$

where the “maximal” sets  $\mathcal{S}_{in_i}$  must also satisfy (11). Therefore, given two maximal sets  $\mathcal{S}_{in_i}, \mathcal{S}_{jn_j}$ , either they must be the same or they must be disjoint, because it is not possible for one of them to be a subset of the other. This shows that the maximal sets  $\mathcal{S}_{in_i}$  are pairwise disjoint once the duplicates are removed, and together span the total feature set  $\mathcal{N} = \{1, \dots, n\}$ . Thus, in a collection of tree-structured sets, the highest level sets do not overlap! Let  $g$  denote the number of distinct maximal sets, and define  $G_i = \mathcal{S}_{in_i}$  for  $i = 1, \dots, g$ . Then  $\{G_1, \dots, G_g\}$  is a partition of  $\mathcal{N}$ , and each  $\mathcal{N}_j$  is a subset of some  $G_i$ . Now define a norm  $\|\cdot\|_{G_j}, j = 1, \dots, g$ , by

$$\|z_{G_j}\|_{G_j} = \sum_{\mathcal{N}_i \subseteq G_j} \|z_{\mathcal{N}_i}\|_i.$$

Because each  $\mathcal{N}_j$  can be a subset of only one  $G_i$ , it follows that the above norm is the same as  $\|\cdot\|_A$ . Therefore this norm is of the form (8) and is thus decomposable.  $\square$

Thus to summarize, the group LASSO norm, the sparse group LASSO norm, and the penalty norms defined in [13, 14] are all decomposable.

With this preparation we can define the sparsity indices and optimal decompositions. Given an integer  $k$ , let  $\text{GkS}$  denote the collection of all group  $k$ -sparse subsets of  $\mathcal{N} = \{1, \dots, n\}$ , and define

$$\sigma_{k,\mathcal{G}}(x, \|\cdot\|) := \min_{\Lambda \in \text{GkS}} \|x - x_\Lambda\| = \min_{\Lambda \in \text{GkS}} \|x_{\Lambda_0^c}\| \quad (12)$$

to be the **group  $k$ -sparsity index** of the vector  $x$  with respect to the norm  $\|\cdot\|$  and the group structure  $\mathcal{G}$ . Since the collection of sets  $\text{GkS}$  is finite (though it could be huge), we are justified in writing  $\min$  instead of  $\inf$ . Once we have the definition of the sparsity index, it is natural to define the next notion. Given  $x \in \mathbb{R}^n$ , and a norm  $\|\cdot\|$ , we call  $\{x_{\Lambda_0}, x_{\Lambda_1}, \dots, x_{\Lambda_s}\}$  an **optimal group  $k$ -sparse decomposition** of  $x$  if  $\Lambda_i \in \text{GkS}$  for  $i = 0, \dots, s$ , and in addition

$$\begin{aligned} \|x_{\Lambda_0^c}\| &= \min_{\Lambda \in \text{GkS}} \|x - x_\Lambda\|, \\ \|x_{\Lambda_i^c}\| &= \min_{\Lambda \in \text{GkS}} \left\| x - \sum_{j=0}^{i-1} x_{\Lambda_j} - x_\Lambda \right\|, i = 1, \dots, s. \end{aligned}$$

There are some wrinkles in group sparsity that do not have any analogs in conventional sparsity. Specifically, suppose  $x \in \mathbb{R}^n$  and that  $\{x_{\Lambda_0}, x_{\Lambda_1}, \dots, x_{\Lambda_s}\}$  is an optimal  $k$ -sparse (*not* optimal *group*  $k$ -sparse) decomposition of  $x$  with respect to  $\|\cdot\|_1$ . Then  $x_{\Lambda_0}$  consists of the  $k$  largest components of  $x$  by magnitude,  $x_{\Lambda_1}$  consists of the next  $k$  largest, and so on. One consequence of this is that

$$\min_j |(x_{\Lambda_i})_j| \geq \max_j |(x_{\Lambda_{i+1}})_j|, \forall i.$$

Therefore

$$\|x_{\Lambda_{i+1}}\|_2 \leq \sqrt{k} \|x_{\Lambda_{i+1}}\|_\infty \leq \frac{1}{\sqrt{k}} \|x_{\Lambda_i}\|_1. \quad (13)$$

This is the equation just above [1, Equation(10)]. However, when we take optimal *group*  $k$ -sparse decompositions, this inequality is no longer valid. For example, suppose  $\|\cdot\|_P = \|\cdot\|_1$ , let  $n = 4, g = 2, k = 2$  and

$$G_1 = \{1, 2\}, G_2 = \{3, 4\}, x = [1 \quad 0.1 \quad 0.6 \quad 0.6]^t.$$

Then it is easy to verify that  $s = 2$ , and

$$\begin{aligned} \Lambda_0 &= \{3, 4\} = G_2, \Lambda_1 = \{1, 2\} = G_1, \\ x_{\Lambda_0} &= [0 \quad 0 \quad 0.6 \quad 0.6]^t, x_{\Lambda_1} = [1 \quad 0.1 \quad 0 \quad 0]^t. \end{aligned}$$



Here we see that the largest element of  $x_{\Lambda_1}$  is in fact larger than the smallest element of  $x_{\Lambda_0}$ . However, we do not have the freedom to “swap” these elements as they belong to different sets  $G_i$ . A more elaborate example is the following: Let  $n = 8, g = 4, k = 4$ , and

$$x = [ 0.1 \quad 1 \quad 0.2 \quad 0.3 \quad 0.4 \quad 0.5 \quad 0.4 \quad 0.7 ],$$

$$G_1 = \{1\}, G_2 = \{2, 3, 4\}, G_3 = \{5, 6\}, G_4 = \{7, 8\}.$$

Then

$$\Lambda_0 = G_3 \cup G_4, \Lambda_1 = G_1 \cup G_2.$$

Note that  $x_{G_2}$  has higher  $\ell_1$ -norm than any other  $x_{G_j}$ . However, since  $G_2$  has cardinality 3, it can only be paired with  $G_1$ , and not with  $G_3$  or  $G_4$ , in order that the cardinality of the union remain less than  $k = 4$ . And  $\|x_{G_1 \cup G_2}\|_1 < \|x_{G_3 \cup G_4}\|_1$ . Therefore an optimal group  $k$ -sparse decomposition of  $x$  is  $x_{G_3 \cup G_4}$  followed by  $x_{G_1 \cup G_2}$ .

### 3 Problem Formulation

The general formulation of a compressed sensing algorithm makes use of three distinct norms:

- $\|\cdot\|_A$  is the *approximation norm* that is used to measure the quality of the approximation. Thus, for a vector  $x \in \mathbb{R}^n$ , the quantity  $\sigma_{k,G}(x, \|\cdot\|_A)$  is the sparsity index used throughout. It is assumed that  $\|\cdot\|_A$  is a *decomposable norm*.
- $\|\cdot\|_P$  is the *penalty norm* that is minimized to induce a desired sparsity structure on the solution. It is assumed that  $\|\cdot\|_P$  is  $\gamma$ -*decomposable* for some  $\gamma \in (0, 1]$ .
- $\|\cdot\|_2$ , which is the standard Euclidean or  $\ell_2$ -norm, and is used to constrain the measurement matrix via the group restricted isometry property (GRIP).

The prototypical problem formulation is this: Suppose  $x \in \mathbb{R}^n$  is an unknown vector,  $A \in \mathbb{R}^{m \times n}$  is a measurement matrix,  $y = Ax + \eta$  is a possibly noise-corrupted measurement vector in  $\mathbb{R}^m$ , and  $\eta \in \mathbb{R}^m$  is the measurement error. It is presumed that  $\|\eta\|_2 \leq \epsilon$ , where  $\epsilon$  is a known prior bound. To estimate  $x$  from  $y$ , we solve the following optimization problem:

$$\hat{x} = \underset{z \in \mathbb{R}^n}{\operatorname{argmin}} \|z\|_P \text{ s.t. } \|y - Az\|_2 \leq \epsilon. \quad (14)$$

The penalty norm  $\|\cdot\|_P$  that is minimized in order to determine an approximation to  $x$  need not be the same as the approximation norm  $\|\cdot\|_A$  used to compute the group  $k$ -sparsity index. If  $\|\cdot\|_P$  is the  $\ell_1$ -, group LASSO, or sparse group LASSO norm, then we take  $\|\cdot\|_A = \|\cdot\|_P$ . The objective is to determine error bounds of the form<sup>2</sup>

$$\|\hat{x} - x\|_2 \leq D_1 \sigma_{k,\mathcal{G}}(x, \|\cdot\|_A) + D_2 \epsilon, \quad (15)$$

or of the form

$$\|\hat{x} - x\|_A \leq D_3 \sigma_{k,\mathcal{G}}(x, \|\cdot\|_A) + D_4 \epsilon, \quad (16)$$

for some appropriate constants  $D_1$  through  $D_4$ .

The interpretation of the inequality (15) in this general setting is the same as in [1, 2]. Suppose the vector  $x$  is group  $k$ -sparse, so that  $\sigma_{k,\mathcal{G}}(x, \|\cdot\|_A) = 0$ . Then an “oracle” that knows the actual support set of  $x$  can approximate  $x$  through computing a generalized inverse of the columns of  $A$  corresponding to the support of  $x$ , and the resulting residual error will be bounded by a multiple of  $\epsilon$ . Now suppose the algorithm satisfies (15). Then (15) implies that the residual error achieved by the algorithm is bounded by a universal constant times that achieved by an oracle. Proceeding further, (15) also implies that if measurements are noise-free so that  $\epsilon = 0$ , then the estimate  $\hat{x}$  equals  $x$ . In other words, the algorithm achieves exact recovery of group  $k$ -sparse vectors under noise-free measurements. Similar remarks apply to the interpretation of the bound (16).

Throughout the paper, we shall be making use of four constants:

$$a := \min_{\Lambda \in \text{GkS}} \min_{x_\Lambda \neq 0} \frac{\|x\|_P}{\|x\|_A}, b := \max_{\Lambda \in \text{GkS}} \max_{x_\Lambda \neq 0} \frac{\|x_\Lambda\|_P}{\|x_\Lambda\|_A}, \quad (17)$$

$$c := \min_{\Lambda \in \text{GkS}} \min_{x_\Lambda \neq 0} \frac{\|x_\Lambda\|_A}{\|x_\Lambda\|_2}, d := \max_{\Lambda \in \text{GkS}} \max_{x_\Lambda \neq 0} \frac{\|x_\Lambda\|_A}{\|x_\Lambda\|_2}. \quad (18)$$

Note that these constants depend on the sparsity structure being used. For instance, in conventional sparsity, as shown below,  $a = b = c = 1$  and  $d = \sqrt{k}$ . The factor  $\sqrt{k}$  is ubiquitous in conventional sparsity, and as shown below, this is where it comes from.

Suppose for instance that  $\|\cdot\|_A = \|\cdot\|_P = \|\cdot\|_1$ , which is the approximation as well as penalty norm used in LASSO. Therefore  $a = b = 1$ . Since

---

<sup>2</sup>The symbol  $A$  is unfortunately doing double duty, representing the approximation norm as well as the measurement matrix. After contemplating various options, it was decided to stick to this notation, in the hope that the context would make clear which usage is meant.

$|\Lambda| \leq k$  for all  $\Lambda \in \text{GkS}$ , we have by Schwarz's inequality that

$$\|v\|_1 \leq \sqrt{k} \|v\|_2$$

whenever  $\text{supp}(v) \subseteq \Lambda \in \text{GkS}$ . In the other direction, we can write

$$v = \sum_{i \in \text{supp}(v)} v_i \mathbf{e}_i,$$

where  $\mathbf{e}_i$  is the  $i$ -th unit vector. Therefore by the triangle inequality

$$\|v\|_2 \leq \sum_{i \in \text{supp}(v)} \|v_i \mathbf{e}_i\|_2 \leq \sum_{i \in \text{supp}(v)} |v_i| = \|v\|_1,$$

and these bounds are tight. Therefore

$$1 = c \leq d = \sqrt{k}.$$

Estimates of these constants for other sparsity-inducing norms are given in Section 5.

## 4 Main Results

In this section we present the main results of the paper. The following definition of the restricted isometry property (RIP) is introduced in [5].

**Definition 4.** Suppose  $A \in \mathbb{R}^{m \times n}$ . Then we say that  $A$  satisfies the **Restricted Isometry Property (RIP)** of order  $k$  with constant  $\delta_k$  if

$$(1 - \delta_k) \|u\|_2^2 \leq \langle u, Au \rangle \leq (1 + \delta_k) \|u\|_2^2, \quad \forall u \in \Sigma_k, \quad (19)$$

where  $\Sigma_k$  denotes the set of all  $u \in \mathbb{R}^n$  such that  $|\text{supp}(u)| \leq k$ .

The first step is to extend the notion of the restricted isometry property (RIP) to the group restricted isometry property (GRIP).

**Definition 5.** A matrix  $A \in \mathbb{R}^{m \times n}$  is said to satisfy the **group restricted isometry property (GRIP)** of order  $k$  with constant  $\delta_k \in (0, 1)$  if

$$1 - \delta_k \leq \min_{\Lambda \in \text{GkS}} \min_{\text{supp}(z) \subseteq \Lambda} \frac{\|Az\|_2^2}{\|z\|_2^2} \leq \max_{\Lambda \in \text{GkS}} \max_{\text{supp}(z) \subseteq \Lambda} \frac{\|Az\|_2^2}{\|z\|_2^2} \leq 1 + \delta_k. \quad (20)$$

Definition 5 shows that the group RIP constant  $\delta_k$  can be smaller than the standard RIP constant in Definition 4, because the various maxima and minima are taken over only group  $k$ -sparse sets, and not all subsets of  $\mathcal{N}$  of cardinality  $k$ . Probabilistic methods for constructing a measurement matrix  $A \in \mathbb{R}^{m \times n}$  that satisfies GRIP with specified order  $k$  and constant  $\delta$  are discussed in Section 6. It is shown that GRIP can be achieved with a smaller value of  $m$  than RIP.

In order to state the main results, we introduce a technical lemma.

**Lemma 2.** *Suppose  $h \in \mathbb{R}^n$ , that  $\Lambda_0 \in GkS$  is arbitrary, and let  $h_{\Lambda_1}, \dots, h_{\Lambda_s}$  be an optimal group  $k$ -sparse decomposition of  $h_{\Lambda_0^c}$  with respect to the decomposable approximation norm  $\|\cdot\|_A$ . Then there exists a constant  $f$  such that*

$$\sum_{j=2}^s \|h_{\Lambda_j}\|_2 \leq \frac{1}{f} \|h_{\Lambda_0^c}\|_A. \quad (21)$$

**Proof:** It is already shown in [1, Equation (11)], [2, Lemma A.4] that

$$\sum_{j=2}^s \|h_{\Lambda_j}\|_2 \leq \frac{1}{\sqrt{k}} \|h_{\Lambda_0^c}\|_1.$$

Therefore, in the case of conventional sparsity, where  $\|\cdot\|_A = \|\cdot\|_P = \|\cdot\|_1$ , one can take  $f = \sqrt{k}$ . In the case of group sparsity, it follows from the definition of the constant  $c$  in (18) that

$$\|h_{\Lambda_j}\|_2 \leq \frac{1}{c} \|h_{\Lambda_j}\|_A, j = 2, \dots, s.$$

Therefore

$$\sum_{j=2}^s \|h_{\Lambda_j}\|_2 \leq \frac{1}{c} \sum_{j=2}^s \|h_{\Lambda_j}\|_A \leq \sum_{j=1}^s \|h_{\Lambda_j}\|_A = \frac{1}{c} \|h_{\Lambda_0^c}\|_A,$$

where the last step follows from the decomposability of  $\|\cdot\|_A$ .  $\square$

Now we state the main theorem for the general optimization problem as stated in (14), and several corollaries for conventional sparsity, group LASSO, sparse group LASSO minimization. All of these are stated at once, followed by a general discussion.

**Theorem 1.** *Suppose that*

1. *The norm  $\|\cdot\|_A$  is decomposable.*

2. The norm  $\|\cdot\|_P$  is  $\gamma$ -decomposable for some  $\gamma \in (0, 1]$ .
3. The matrix  $A$  satisfies GRIP of order  $2k$  with constant  $\delta_{2k}$ .
4. Suppose the “compressibility condition”

$$\delta_{2k} < \frac{fa\gamma}{\sqrt{2} + fa\gamma/bd} \quad (22)$$

holds, where  $d$  is defined in (18) and  $f$  is defined in Lemma 2.

Define

$$\hat{x} = \operatorname{argmin}_{z \in \mathbb{R}^n} \|z\|_P \text{ s.t. } \|y - Az\|_2 \leq \epsilon. \quad (23)$$

Then

$$\|\hat{x} - x\|_2 \leq D_1 \sigma_{k,G}(x, \|\cdot\|_A) + D_2 \epsilon, \quad (24)$$

where

$$D_1 = \frac{r(1+\gamma)}{f} \cdot \frac{1 + (\sqrt{2} - 1)\delta_{2k}}{1 - (1 + \sqrt{2}rd/f)\delta_{2k}}, \quad (25)$$

$$D_2 = 2(1 + rd/f) \frac{\sqrt{1 + \delta_{2k}}}{1 - (1 + \sqrt{2}rd/f)\delta_{2k}}. \quad (26)$$

Further,

$$\|\hat{x} - x\|_A \leq D_3 \sigma_{k,G}(x, \|\cdot\|_A) + D_4 \epsilon, \quad (27)$$

where

$$D_3 = r(1+\gamma) \cdot \frac{1 + (\sqrt{2}d/f - 1)\delta_{2k}}{1 - (1 + \sqrt{2}rd/f)\delta_{2k}}, \quad (28)$$

$$D_4 = 2(1 + rd) \frac{\sqrt{1 + \delta_{2k}}}{1 - (1 + \sqrt{2}rd/f)\delta_{2k}}. \quad (29)$$

**Corollary 4. (Conventional Sparsity)** Define

$$\hat{x}_{\text{CS}} = \operatorname{argmin}_z \|z\|_1 \text{ s.t. } \|y - Az\|_2 \leq \epsilon. \quad (30)$$

Then Theorem 1 applies with  $\|\cdot\|_A = \|\cdot\|_P = \|\cdot\|_1$ ,

$$a = 1, b = 1, c = 1, d = \sqrt{k}, f = \sqrt{k}, \gamma = 1. \quad (31)$$

Therefore the compressibility condition (22) becomes

$$\delta_{2k} < \sqrt{2} - 1. \quad (32)$$

This leads to the error bounds

$$\|\hat{x}_{\text{CS}} - x\|_2 \leq D_2 \sigma_k(x, \|\cdot\|_1) + D_2 \epsilon, \quad (33)$$

$$\|\hat{x}_{\text{CS}} - x\|_1 \leq D_3 \sigma_k(x, \|\cdot\|_1) + D_4 \epsilon, \quad (34)$$

where

$$D_1 = \frac{2}{\sqrt{k}} \frac{1 + (\sqrt{2} - 1)\delta_{2k}}{1 - (1 + \sqrt{2}\delta_{2k})}, D_2 = 4 \frac{1 + \sqrt{\delta_{2k}}}{1 - (1 + \sqrt{2}\delta_{2k})},$$

$$D_3 = 2 \frac{1 + (\sqrt{2} - 1)\delta_{2k}}{1 - (1 + \sqrt{2}\delta_{2k})} \sigma_k(x, \|\cdot\|_1), D_4 = 4\sqrt{k} \frac{1 + \sqrt{\delta_{2k}}}{1 - (1 + \sqrt{2}\delta_{2k})}.$$

**Corollary 5. (Group LASSO)** Suppose  $\{G_1, \dots, G_g\}$  is a partition of  $\mathcal{N} = \{1, \dots, n\}$ , and that  $l_{\min} \leq |G_j| \leq k$  for all  $j$ . Let  $s_{\max} = \lfloor k/l_{\min} \rfloor$ , and define the group LASSO norm

$$\|z\|_{\text{GL}} = \sum_{j=1}^g \|z_{G_j}\|_2. \quad (35)$$

Define the estimate

$$\hat{x}_{\text{GL}} = \underset{z}{\operatorname{argmin}} \|z\|_{\text{GL}} \text{ s.t. } \|y - Az\|_2 \leq \epsilon. \quad (36)$$

Then Theorem 1 applies with  $\|\cdot\|_A = \|\cdot\|_P = \|\cdot\|_{\text{GL}}$ ,

$$a = 1, b = 1, c = 1, d = \sqrt{s_{\max}}, f = 1, \gamma = 1. \quad (37)$$

Therefore the compressibility condition (22) becomes

$$\delta_{2k} < \frac{1}{\sqrt{2s_{\max}} + 1} \quad (38)$$

This leads to the error bounds

$$\|\hat{x}_{\text{GL}} - x\|_2 \leq D_1 \sigma_k(x, \|\cdot\|_{\text{GL}}) + D_2 \epsilon,$$

and

$$\|\hat{x}_{\text{GL}} - x\|_{\text{GL}} \leq D_3 \sigma_k(x, \|\cdot\|_{\text{GL}}) + D_4 \epsilon,$$

where

$$D_1 = 2 \frac{1 + (\sqrt{2} - 1)\delta_{2k}}{1 - (1 + \sqrt{2s_{\max}})\delta_{2k}}, D_2 = 4 \frac{1 + \sqrt{\delta_{2k}}}{1 - (1 + \sqrt{2s_{\max}})\delta_{2k}},$$

$$D_3 = 2 \frac{1 + (\sqrt{2} - 1)\delta_{2k}}{1 - (1 + \sqrt{2s_{\max}})\delta_{2k}}, D_4 = 4\sqrt{s_{\max}} \frac{1 + \sqrt{\delta_{2k}}}{1 - (1 + \sqrt{2s_{\max}})\delta_{2k}}.$$

**Corollary 6. (Sparse Group LASSO)** Suppose  $\{G_1, \dots, G_g\}$  is a partition of  $\mathcal{N} = \{1, \dots, n\}$ , and that  $l_{\min} \leq |G_j| \leq l_{\max}$  for all  $j$ . Let  $s_{\max} = \lfloor k/l_{\min} \rfloor$ , and define the sparse group LASSO norm

$$\|z\|_{\text{SGL}, \mu} = \sum_{j=1}^g [(1 - \mu)\|z_{G_j}\|_1 + \mu\|z_{G_j}\|_2]. \quad (39)$$

Define the estimate

$$\hat{x}_{\text{SGL}} = \underset{z}{\operatorname{argmin}} \|z\|_{\text{SGL}, \mu} \text{ s.t. } \|y - Az\|_2 \leq \epsilon. \quad (40)$$

Then Theorem 1 applies with  $\|\cdot\|_A = \|\cdot\|_P = \|\cdot\|_{\text{SGL}}$ ,

$$a = 1, b = 1, c = 1, d = (1 - \mu)\sqrt{l_{\max}} + \mu\sqrt{s_{\max}}, f = 1, \gamma = 1. \quad (41)$$

Therefore the compressibility condition (22) becomes

$$\delta_{2k} < \frac{d}{\sqrt{2} + d}, \quad (42)$$

where  $d$  is defined in (41). This leads to the error bounds

$$\|\hat{x}_{\text{SGL}} - x\|_2 \leq D_1 \sigma_k(x, \|\cdot\|_{\text{SGL}}) + D_2 \epsilon,$$

and

$$\|\hat{x}_{\text{SGL}} - x\|_{\text{SGL}} \leq D_3 \sigma_k(x, \|\cdot\|_{\text{GL}}) + D_4 \epsilon,$$

where the constants  $D_1$  through  $D_4$  are the same as in Corollary 5 with the term  $\sqrt{s_{\max}}$  replaced by  $d$  as shown in (41).

Before presenting the proofs of these bounds, we briefly discuss their implications.

1. In the case of conventional sparsity, the bounds on  $\|\hat{x} - x\|_2$  and  $\|\hat{x} - x\|_1$  reduce to those proved earlier in [1, 2, 3]. To the best of the authors' knowledge, there are no bounds of the form (2) and (3) available for other penalty norms. Therefore the bounds in Theorem 1 contain known bounds as special cases and some new bounds as well.
2. In the case of conventional sparsity, the upper bound on  $\|\hat{x} - x\|_1$  is precisely  $\sqrt{k}$  times the upper bound on  $\|\hat{x} - x\|_2$ . Note that if the vector  $\hat{x} - x$  is  $k$ -sparse, then by Schwarz' inequality it would follow that  $\|\hat{x} - x\|_1 \leq \sqrt{k}\|\hat{x} - x\|_2$ . It is therefore interesting that a similar relationship holds even though the residual error  $\hat{x} - x$  need not be  $k$ -sparse.

3. In the case of the group LASSO norm, the key parameter is  $s_{\max}$ , the largest number of sets  $G_i$  that can comprise any group  $k$ -sparse set. If each set  $G_i$  is a singleton, then  $s_{\max} = k$ .
4. The only difference between the bounds for the group LASSO and the sparse group LASSO norms is in the parameter  $d$ .

## 5 Proofs of Main Results

The proof of Theorem 1 depends on a few preliminary lemmas.

Lemma 3 should be compared with [1, Lemma 2.1], [2, Lemma A.3].

**Lemma 3.** *Suppose  $A \in \mathbb{R}^{m \times n}$  satisfies the group RIP of order  $2k$  with constant  $\delta_{2k}$ , and that  $u, v$  are group  $k$ -sparse with supports contained in disjoint group  $k$ -sparse subsets of  $\mathcal{N}$ . Then*

$$|\langle Au, Av \rangle| \leq \delta_{2k} \|u\|_2 \cdot \|v\|_2. \quad (43)$$

**Proof:** Since we can divide through by  $\|u\|_2 \cdot \|v\|_2$ , an equivalent statement is the following: If  $u, v$  are group  $k$ -sparse with supports contained in disjoint group  $k$ -sparse subsets of  $\mathcal{N}$ , and  $\|u\|_2 = \|v\|_2 = 1$ , then

$$|\langle Au, Av \rangle| \leq \delta_{2k}.$$

Now the assumptions guarantee that  $u \pm v$  are both group  $2k$ -sparse. Moreover  $u^t v = 0$  since they have disjoint support. Therefore  $\|u \pm v\|_2^2 = 2$ . So the group RIP implies that

$$2(1 - \delta_{2k}) \leq \|Au \pm Av\|_2^2 \leq 2(1 + \delta_{2k}).$$

Now the parallelogram identity implies that

$$|\langle Au, Av \rangle| = \left| \frac{\|Au + Av\|_2^2 - \|Au - Av\|_2^2}{4} \right| \leq \delta_{2k}.$$

This is the desired conclusion.  $\square$

**Lemma 4.** *Suppose  $h \in \mathbb{R}^n$ , that  $\Lambda_0 \in GkS$  is arbitrary, and let  $h_{\Lambda_1}, \dots, h_{\Lambda_s}$  be an optimal group  $k$ -sparse decomposition of  $h_{\Lambda_0^c}$  with respect to the approximation norm  $\|\cdot\|_A$ . Define  $\Lambda = \Lambda_0 \cup \Lambda_1$ . Then*

$$\|h_\Lambda\|_2 \leq \frac{\sqrt{2}\delta_{2k}}{f(1 - \delta_{2k})} \|h_{\Lambda_0^c}\|_A + \frac{\sqrt{(1 + \delta_{2k})}}{(1 - \delta_{2k})} \|Ah\|_2. \quad (44)$$



The proof closely mimics that of [2, Lemma 1.3]. But it is presented in detail, in the interests of completeness.

**Proof:** Note that  $h_\Lambda$  is group  $2k$ -sparse. Therefore by the definition of the group RIP property, it follows that

$$(1 - \delta_{2k})\|h_\Lambda\|_2^2 \leq \|Ah_\Lambda\|_2^2 \leq (1 + \delta_{2k})\|h_\Lambda\|_2^2.$$

Next, observe that

$$\|Ah_\Lambda\|_2^2 = \langle Ah_\Lambda, Ah_\Lambda \rangle.$$

So we will work on a bound for the right side. Note that

$$\langle Ah_\Lambda, Ah_\Lambda \rangle = \langle Ah_\Lambda, Ah \rangle - \langle Ah_\Lambda, Ah_{\Lambda^c} \rangle.$$

Next by (27) and Schwarz's inequality, it follows that

$$\begin{aligned} |\langle Ah_\Lambda, Ah_{\Lambda^c} \rangle| &\leq \left| \sum_{i=0}^1 \sum_{j=2}^s \langle Ah_{\Lambda_i}, Ah_{\Lambda_j} \rangle \right| \\ &\leq \delta_{2k} [\|h_{\Lambda_0}\|_2 + \|h_{\Lambda_1}\|_2] \sum_{j=2}^s \|h_{\Lambda_j}\|_2 \\ &\leq \frac{\sqrt{2}\delta_{2k}}{f} \|h_\Lambda\|_2 \|h_{\Lambda^c}\|_A. \end{aligned}$$

In the above, we use the known inequality

$$\|h_{\Lambda_0}\|_2 + \|h_{\Lambda_1}\|_2 \leq \sqrt{2}\|h_{\Lambda_0} + h_{\Lambda_1}\|_2 = \sqrt{2}\|h_\Lambda\|_2,$$

because  $h_{\Lambda_0}$  and  $h_{\Lambda_1}$  are orthogonal. Next

$$|\langle Ah_\Lambda, Ah \rangle| \leq \|Ah_\Lambda\|_2 \cdot \|Ah\|_2 \leq \sqrt{(1 + \delta_{2k})} \|h_{\Lambda_0}\|_2 \cdot \|Ah\|_2.$$

Combining everything gives

$$\begin{aligned} (1 - \delta_{2k})\|h_\Lambda\|_2^2 &\leq \|Ah_\Lambda\|_2^2 \\ &\leq |\langle Ah_\Lambda, Ah \rangle| + |\langle Ah_\Lambda, Ah_{\Lambda^c} \rangle| \\ &\leq \frac{\sqrt{2}\delta_{2k}}{f} \|h_\Lambda\|_2 \|h_{\Lambda^c}\|_A + \sqrt{(1 + \delta_{2k})} \|h_{\Lambda_0}\|_2 \cdot \|Ah\|_2. \end{aligned}$$

Dividing both sides by  $(1 - \delta_{2k})\|h_\Lambda\|_2$  leads to (44).  $\square$

**Proof of Theorem 1:** Define  $\hat{x}$  as in (23), and define  $h = \hat{x} - x$ , so that  $\hat{x} = x + h$ . The optimality of  $\hat{x}$  implies that  $\|x\|_P \geq \|\hat{x}\|_P = \|x + h\|_P$ .

Let  $\{x_{\Lambda_0}, x_{\Lambda_1}, \dots, x_{\Lambda_s}\}$  be an optimal group  $k$ -sparse decomposition of  $x$ . Then the triangle inequality and the optimality of  $\hat{x}$  together imply that

$$\sum_{i=0}^s \|x_{\Lambda_i}\|_P \geq \|x\|_P \geq \|x + h\|_P. \quad (45)$$

Now the  $\gamma$ -decomposability of  $\|\cdot\|_P$  implies that

$$\begin{aligned} \|x + h\|_P &\geq \|x_{\Lambda_0} + h_{\Lambda_0} + x_{\Lambda_0^c} + h_{\Lambda_0^c}\|_P \\ &\geq \|x_{\Lambda_0} + h_{\Lambda_0}\|_P + \gamma \sum_{i=1}^s \|x_{\Lambda_i} + h_{\Lambda_i}\|_P \\ &\geq \|x_{\Lambda_0}\|_P - \|h_{\Lambda_0}\|_P + \gamma \sum_{i=1}^s [\|h_{\Lambda_i}\|_P - \|x_{\Lambda_i}\|_P]. \end{aligned} \quad (46)$$

Combining (45) and (46), cancelling the common term  $\|x_{\Lambda_0}\|_P$ , and rearranging leads to

$$\gamma \sum_{i=1}^s \|h_{\Lambda_i}\|_P \leq \|h_{\Lambda_0}\|_P + (1 + \gamma) \sum_{i=1}^s \|x_{\Lambda_i}\|_P.$$

Next we make use the definition of the constants  $a$  and  $b$  from (17), the decomposability of  $\|\cdot\|_A$ , and the triangle inequality. This leads to

$$\begin{aligned} a\gamma \|h_{\Lambda_0^c}\|_A &= a\gamma \sum_{i=1}^s \|h_{\Lambda_i}\|_A \\ &\leq \gamma \sum_{i=1}^s \|h_{\Lambda_i}\|_P \\ &\leq \|h_{\Lambda_0}\|_P + (1 + \gamma) \sum_{i=1}^s \|x_{\Lambda_i}\|_P \\ &\leq b\|h_{\Lambda_0}\|_A + b(1 + \gamma) \sum_{i=1}^s \|x_{\Lambda_i}\|_A \\ &= b\|h_{\Lambda_0}\|_A + b(1 + \gamma) \|x_{\Lambda_0^c}\|_A \\ &= b\|h_{\Lambda_0}\|_A + b(1 + \gamma) \sigma_A, \end{aligned} \quad (47)$$

where  $\sigma_A$  is shorthand for  $\sigma_{k, \mathcal{G}}(x, \|\cdot\|_A)$ , the group  $k$ -sparsity index of  $x$ . Dividing both sides by  $a\gamma$  gives

$$\|h_{\Lambda_0^c}\|_A \leq r\|h_{\Lambda_0}\|_A + r(1 + \gamma) \sigma_A, \quad (48)$$

where  $r = b/a\gamma$ . Next, it follows from the definition of  $d$  in (18) that

$$\|h_{\Lambda_0}\|_A \leq d\|h_{\Lambda_0}\|_2 \leq d\|h_\Lambda\|_2,$$

where as before  $\Lambda = \Lambda_0 \cup \Lambda_1$ . Substituting into the previous bound gives

$$\|h_{\Lambda_0^c}\|_A \leq rd\|h_\Lambda\|_2 + r(1 + \gamma)\sigma_A. \quad (49)$$

This is the first of two inequalities that we require.

Next, both  $x$  and  $\hat{x}$  are feasible for the optimization problem in (23). This implies that

$$\|Ah\|_2 \leq \|A(\hat{x} - x)\|_2 \leq \|A\hat{x} - y\|_2 + \|Ax - y\|_2 \leq 2\epsilon.$$

Therefore (44) now becomes

$$\|h_\Lambda\|_2 \leq \frac{\sqrt{2}\delta_{2k}}{f(1 - \delta_{2k})}\|h_{\Lambda_0^c}\|_A + \frac{2\sqrt{1 + \delta_{2k}}}{(1 - \delta_{2k})}\epsilon. \quad (50)$$

Define the symbols

$$g = \frac{\sqrt{2}\delta_{2k}}{(1 - \delta_{2k})}, r_2 = \frac{2\sqrt{1 + \delta_{2k}}}{(1 - \delta_{2k})}, \quad (51)$$

so that  $g$  and  $r_2$  depend only the GRIP constant  $\delta_{2k}$ . Therefore (52) can be expressed compactly as

$$\|h_\Lambda\|_2 \leq (g/f)\|h_{\Lambda_0^c}\|_A + r_2\epsilon. \quad (52)$$

This is the second inequality we require.

The inequalities (49) and (52) can be written as a vector inequality, namely

$$\begin{bmatrix} 1 & -rd \\ -g/f & 1 \end{bmatrix} \begin{bmatrix} \|h_{\Lambda_0^c}\|_A \\ \|h_\Lambda\|_2 \end{bmatrix} \leq \begin{bmatrix} r(1 + \gamma) \\ 0 \end{bmatrix} \sigma_A + \begin{bmatrix} 0 \\ r_2 \end{bmatrix} \epsilon.$$

The coefficient matrix on the left side has a strictly positive inverse if its determinant  $1 - grd/f$  is positive. So the ‘‘compressibility condition’’ is  $g < f/rd$ , which is the same as (22). Moreover, if  $1 - grd/f > 0$ , then one can infer from the above vector inequality that

$$\begin{aligned} \begin{bmatrix} \|h_{\Lambda_0^c}\|_A \\ \|h_\Lambda\|_2 \end{bmatrix} &\leq \frac{1}{1 - grd/f} \begin{bmatrix} 1 & rd \\ g/f & 1 \end{bmatrix} \left\{ \begin{bmatrix} r(1 + \gamma) \\ 0 \end{bmatrix} \sigma_A + \begin{bmatrix} 0 \\ r_2 \end{bmatrix} \epsilon \right\} \\ &= \frac{1}{1 - grd/f} \left\{ \begin{bmatrix} 1 \\ g/f \end{bmatrix} r(1 + \gamma)\sigma_A + \begin{bmatrix} rd \\ 1 \end{bmatrix} r_2\epsilon \right\}. \end{aligned}$$

Now by (21),

$$\|h_{\Lambda^c}\|_2 \leq \sum_{j=2}^s \|h_{\Lambda_j}\|_2 \leq \frac{1}{f} \|h_{\Lambda_0^c}\|_A.$$

Therefore, since  $h = h_{\Lambda} + h_{\Lambda^c}$ , the triangle inequality implies that

$$\begin{aligned} \|h\|_2 &\leq \|h_{\Lambda^c}\|_2 + \|h_{\Lambda}\|_2 \\ &\leq \frac{1}{f} \|h_{\Lambda_0^c}\|_A + \|h_{\Lambda}\|_2 \\ &\leq \frac{1}{1 - grd/f} \begin{bmatrix} 1/f & 1 \end{bmatrix} \left\{ \begin{bmatrix} 1 \\ g/f \end{bmatrix} r(1 + \gamma)\sigma_A + \begin{bmatrix} rd \\ 1 \end{bmatrix} r_2\epsilon \right\} \\ &= \frac{1}{1 - grd/f} [r(1 + \gamma)(1/f + g/f)\sigma_A + (1 + (rd)/f)r_2\epsilon]. \end{aligned}$$

Substituting for the various constants and clearing leads to the bound in (24).

To derive the bound (27) on  $\|\hat{x} - x\|_A$ , we adopt the same strategy of deriving a vector inequality and then inverting the coefficient matrix. We already have from (48) that

$$\|h_{\Lambda_0^c}\|_A \leq r \|h_{\Lambda_0}\|_A + r(1 + \gamma)\sigma_A.$$

Next, it follows from the definition of  $d$  in (18) and (50) that

$$\|h_{\Lambda_0}\|_A \leq d \|h_{\Lambda_0}\|_2 \leq d \|h_{\Lambda}\|_2 \leq \frac{gd}{f} \|h_{\Lambda_0^c}\|_A + r_2\epsilon,$$

where  $g$  and  $r_2$  are defined in (51). These two inequalities can be combined into the vector inequality

$$\begin{bmatrix} 1 & -r \\ -gd/f & 1 \end{bmatrix} \begin{bmatrix} \|h_{\Lambda_0^c}\|_A \\ \|h_{\Lambda_0}\|_A \end{bmatrix} \leq \begin{bmatrix} r(1 + \gamma)\sigma_A \\ r_2\epsilon \end{bmatrix}.$$

Though the coefficient matrix is different, the determinant is still  $1 - rdg/f$ . Therefore, if (22) holds, then the coefficient matrix has a positive inverse. In this case we can conclude that

$$\begin{aligned} \|h\|_A &\leq \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} \|h_{\Lambda_0^c}\|_A \\ \|h_{\Lambda_0}\|_A \end{bmatrix} \\ &\leq \frac{1}{1 - rdg/f} \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & r \\ gd/f & 1 \end{bmatrix} \begin{bmatrix} r(1 + \gamma)\sigma_A \\ r_2\epsilon \end{bmatrix}. \end{aligned}$$

After clearing terms, this is the bound (27).  $\square$

**Proof of Corollary 4:** If both  $\|\cdot\|_A$  and  $\|\cdot\|_P$  are equal, it is obvious that  $a = b = 1$ , as defined in (17). Next, it is a ready consequence of Schwarz' inequality that  $c = 1$  and  $d = \sqrt{k}$ , as defined in (18). Next, it is shown in [1, Equation (11)], [2, Lemma A.4] that  $f$  defined in (21) equals  $\sqrt{k}$ . Because  $\|\cdot\|_P$  is decomposable, we can take  $\gamma = 1$ . Substituting these values into the bound (24) through (27) establishes the desired bounds (33) and (34).  $\square$

**Proof of Corollary 5:** Let  $\|\cdot\|_A = \|\cdot\|_P = \|\cdot\|_{GL}$ . Then since  $\|\cdot\|_P = \|\cdot\|_A$ , we have  $a = b = 1$ . To calculate  $c$  and  $d$ , define  $l_{\min}$  to be the smallest cardinality of any  $G_i$ , and define  $s_{\max} := \lfloor k/l_{\min} \rfloor$ . Now suppose that  $\Lambda \in \text{GkS}$ . Specifically, suppose  $\Lambda = G_{i_1} \cup \dots \cup G_{i_s}$ . Then clearly

$$\|z_\Lambda\|_{GL} = \sum_{j=1}^s \|z_{G_{i_j}}\|_2,$$

while

$$\|z_\Lambda\|_2 = \left( \sum_{j=1}^s \|z_{G_{i_j}}\|_2^2 \right)^{1/2}.$$

Thus, if we define the  $s$ -dimensional vector  $v \in \mathbb{R}_+^s$  by

$$v = [\|z_{G_{i_j}}\|_2, j = 1, \dots, s],$$

then

$$\|z_\Lambda\|_{GL} = \|v\|_1, \|z_\Lambda\|_2 = \|v\|_2.$$

Now it is easy to see that

$$\|v\|_2 \leq \|v\|_1 \leq \sqrt{s}\|v\|_2.$$

Moreover, it is clear that the integer  $s$ , denoting the number of distinct sets that make up  $\Lambda$ , cannot exceed  $s_{\max}$ . This shows that

$$1 \leq c_{GL} \leq d_{GL} \leq \sqrt{s_{\max}}. \quad (53)$$

As shown in the proof of Lemma 2, in the case of group sparsity, one can only take  $f = c = 1$ . Finally, because  $\|\cdot\|_P$  is decomposable, it follows that  $\gamma = 1$ . Substituting these values into (24) through (29) leads to the desired bounds.  $\square$

**Proof of Corollary 6:** In this case  $\|\cdot\|_A = \|\cdot\|_P = \|\cdot\|_{SGL, \mu}$ . Because both norms are equal, it follows that  $a = b = 1$ . To calculate  $c$  and  $d$ ,

suppose  $\Lambda = G_{i_1} \cup \dots \cup G_{i_s}$ . Let  $l_{\max}$  denote the largest cardinality of any  $G_i$ . Then

$$\|z_{G_{i_j}}\|_2 \leq \|z_{G_{i_j}}\|_1 \leq \sqrt{l_{\max}} \|z_{G_{i_j}}\|_2,$$

whence

$$\sum_{j=1}^s \|z_{G_{i_j}}\|_2 \leq \sum_{j=1}^s \|z_{G_{i_j}}\|_1 \leq \sqrt{l_{\max}} \sum_{j=1}^s \|z_{G_{i_j}}\|_2. \quad (54)$$

Combining (53) and (54) leads to

$$\|z_{\Lambda}\|_2 \leq \|z_{\Lambda}\|_{\text{SGL},\mu} \leq [(1-\mu)\sqrt{l_{\max}} + \mu\sqrt{s_{\max}}] \|z_{\Lambda}\|_2.$$

Therefore

$$1 \leq c_{\text{SGL},\mu} \leq d_{\text{SGL},\mu} \leq (1-\mu)\sqrt{l_{\max}} + \mu\sqrt{s_{\max}}. \quad (55)$$

Again, in the case of group sparsity one has to take  $f = c = 1$ . Finally, because  $\|\cdot\|_P$  is decomposable, we can take  $\gamma = 1$ . Substituting these values into (24) through (29) leads to the desired bounds.  $\square$

## 6 Bounds on the Number of Measurements

In this section we study the following problem: Suppose a matrix  $A \in \mathbb{R}^{m \times n}$  is constructed by drawing  $mn$  i.i.d. samples of a fixed random variable  $X$ . Suppose we are specified integers  $n, k \ll n$ , and real numbers  $\delta, \zeta \in (0, 1)$ . The objective is to determine a lower bound on  $m$  such that  $A$  satisfies GRIP or order  $k$  with constant  $\delta$ , with probability no smaller than  $1 - \zeta$ .

The approach here follows [17, 3]. Recall that a zero-mean random variable  $X$  is said to be **sub-Gaussian** if there exist constants  $\alpha, \beta$  such that

$$\Pr\{|X| > \epsilon\} \leq \alpha \exp(-\beta\epsilon^2), \quad \forall \epsilon > 0. \quad (56)$$

A normal random variable satisfies (56) with  $\alpha = 2, \beta = 0.5$ . Suppose in addition that  $X$  has unit variance, and define  $A \in \mathbb{R}^{m \times n}$  by drawing  $mn$  i.i.d. samples of  $X/m$ . Then it is known ([3, Lemma 9.8]) that

$$\Pr\{|\|Au\|_2^2 - \|u\|_2^2| > \epsilon \|u\|_2^2\} \leq 2 \exp(-m c \epsilon^2),$$

where

$$c = \frac{\beta^2}{4\alpha + 2\beta}. \quad (57)$$

With this background, we can begin to address the problem under study.

**Lemma 5.** *Given integers  $n, k \ll n$  and a real number  $\delta \in (0, 1)$ , and any collection  $\mathcal{J}$  of subsets of  $\mathcal{N} = \{1, \dots, n\}$  such that  $|T| \leq k \forall T \in \mathcal{J}$ . Let  $X$  be a zero-mean, unit variance, sub-Gaussian random variable satisfying (56), and let  $A \in \mathbb{R}^{m \times n}$  consist of  $mn$  i.i.d. samples of  $X$ . Then*

$$(1 - \delta)\|x_T\|_2^2 \leq \|Ax_T\|_2^2 \leq (1 + \delta)\|x_T\|_2^2 \quad \forall T \in \mathcal{J}, \quad \forall x \in \mathbb{R}^n \quad (58)$$

with probability no smaller than  $1 - \zeta$ , where  $\zeta$  is given by

$$\zeta = 2|\mathcal{J}| \left( \frac{12}{\theta} \right)^k \exp(-mc\theta^2), \quad (59)$$

where  $c$  is defined in (57) and

$$\theta = 1 - \sqrt{1 - \delta}. \quad (60)$$

**Proof:** It is shown in [17, Lemma 5.1] that, for a given fixed index set  $T \subseteq \mathcal{N}$  with  $|T| \leq k$ , the inequality

$$(1 - \theta)\|x_T\|_2 \leq \|Ax_T\|_2 \leq (1 + \theta)\|x_T\|_2, \quad \forall x \in \mathbb{R}^n, \quad (61)$$

with probability no smaller than  $1 - \zeta'$ , where

$$\zeta' = 2 \left( \frac{12}{\theta} \right)^k \exp(-mc\theta^2). \quad (62)$$

However, the inequality (61) does not quite match the definition of RIP or GRIP, because the inequality involves  $\|Ax_T\|_2$  and not  $\|Ax_T\|_2^2$ . Therefore, in order to convert (62) into (58), we need to have

$$1 - \delta \leq (1 - \theta)^2, \text{ and } (1 + \theta)^2 \leq 1 + \delta,$$

or equivalently,

$$\theta \leq \max\{1 - \sqrt{1 - \delta}, \sqrt{1 + \delta} - 1\}.$$

It is elementary to show that the first term is always larger than the second, so that (61) implies (58) provided  $\theta$  is defined as in (60).

Next, suppose the collection  $\mathcal{J}$  is specified. Then [17, Lemma 5.1] implies that (61) holds for each fixed set with probability no smaller than  $1 - \zeta'$ . Therefore the union of events bound shows that (58) holds with probability no smaller than  $1 - |\mathcal{J}|\zeta'$ , where  $\zeta'$  is defined in (62). The proof is completed by noting that  $\zeta$  defined in (59) is precisely  $|\mathcal{J}|\zeta'$ .  $\square$

Now we are ready to give estimates for the integer  $m$ .

**Theorem 2.** Suppose integers  $n, k \ll n$  are specified, together with real numbers  $\delta, \zeta \in (0, 1)$ . Let  $X$  be a sub-Gaussian zero-mean unit-variance random variable, and define the constant  $c$  as in (57). Let  $A \in \mathbb{R}^{m \times n}$  consist of  $mn$  i.i.d. random samples of  $X/m$ . Define  $\theta$  as in (60). Then

1.  $A$  satisfies RIP of order  $k$  with constant  $\delta$ , with probability no smaller than  $1 - \zeta$ , provided

$$m_S \geq \frac{1}{c\theta^2} \left[ \log \frac{2}{\zeta} + k \left( \log \frac{en}{k} + \log \frac{12}{\theta} \right) \right]. \quad (63)$$

2. Suppose  $\{G_1, \dots, G_g\}$  is a partition of  $\mathcal{N} = \{1, \dots, n\}$ , where  $l_{\min} \leq |G_i| \leq k$  for all  $i$ . Define  $s_{\max} = \lfloor k/l_{\min} \rfloor$ . Then  $A$  satisfies GRIP of order  $k$  with constant  $\delta$ , with probability no smaller than  $1 - \zeta$ , provided

$$m_{GS} \geq \frac{1}{c\theta^2} \left[ \log \frac{2}{\zeta} + s_{\max} \log \frac{eg}{s_{\max}} + k \log \frac{12}{\theta} \right]. \quad (64)$$

**Proof:** Suppose a set  $S$  consists of  $s$  elements, and that  $t < s$ . Then the number of distinct subsets of  $S$  with  $t$  or fewer elements is given by

$$\sum_{i=0}^t \binom{s}{i} \leq \left( \frac{es}{t} \right)^t,$$

where the bound is a part of Sauer's lemma, which can be found in many places, out of which [18, Theorem 4.1] is just one reference. To prove (1), note that the number of distinct subsets of  $\mathcal{N}$  with  $k$  or fewer elements is bounded by  $(en/k)^k$  by Sauer's lemma. Therefore, given  $n, k, \delta, \zeta$ , one can choose  $m$  large enough that

$$2 \left( \frac{en}{k} \right)^k \left( \frac{12}{\theta} \right)^k \exp(-mc\theta^2) \leq \zeta,$$

which is equivalent to (63), and  $A$  would satisfy RIP of order  $k$  with constant  $\delta$  with probability no less than  $1 - \zeta$ . To prove Item 2, note that every group  $k$ -sparse set is a union of at most  $s_{\max}$  sets among  $G_1, \dots, G_g$ . Therefore the number of group  $k$ -sparse subsets of  $\mathcal{N}$  is bounded by  $(eg/s_{\max})^{s_{\max}}$ . Therefore, given  $n, k, \delta, \zeta$ , one can choose  $m$  large enough that

$$2 \left( \frac{eg}{s_{\max}} \right)^{s_{\max}} \left( \frac{12}{\theta} \right)^k \exp(-mc\theta^2) \leq \zeta,$$



which is equivalent to (64), and  $A$  would satisfy GRIP or order  $k$  with constant  $\delta$  with probability no less than  $1 - \zeta$ .  $\square$

One of the nice features of these bounds (62) and (63) is that in both cases the confidence level  $\zeta$  enters through the logarithm, so that  $m$  increases very slowly as we decrease  $\zeta$ . This is consistent with the well-known maxim in statistical learning theory that “confidence is cheaper than accuracy.”

Next we compare the number of measurements required with conventional versus group sparsity. It is pointed out in [10] that if random projections are used to construct  $A$ , then satisfying the group RIP requires fewer samples than satisfying RIP. In particular, suppose all groups have the same size  $s$ , implying that  $n = gs$  where  $g$  is the number of groups. Suppose also that  $k$  is a multiple of  $s$ , say  $k = sr$ . Then satisfying the group RIP condition requires only  $O(k + r \log g)$  random projections, whereas satisfying the RIP requires  $O(k \log n)$  random projections. The bounds in Theorem 2 generalize these observations, as they do not require that all groups must be of the same size, or that either  $n$  or  $k$  be a multiple of the group size. Note that, when  $\delta$  is very small,  $\theta \approx \delta/2$ . Therefore a comparison of (62) and (63) shows that  $m_S$  is  $O(k \log n)/\delta^2$ , whereas  $m_{GS}$  is  $O(k + s_{\max} \log g)/\delta^2$ . This is the generalization of the term involving  $s_{\max}$  will dominate the term involving  $k$ . So in principle group sparsity would require fewer measurements than conventional sparsity. However, since  $s_{\max}$  is multiplied by  $\log g$ ,  $g$  would have to be truly enormous in order for group sparsity to lead to substantially smaller values for  $m$  than conventional sparsity.

The important point is that, unless  $n$  is extremely large, neither of the bounds (62) or (63) leads to a value of  $m$  that is smaller than  $n$ . To illustrate this last comment, let us apply the bounds from Theorem 2 to typical numbers from microarray experiments in cancer biology. Accordingly, we take  $n = 20,000$ , which is roughly equal to the number of genes in the human body and the number of measured quantities in a typical experiment, and we take  $k = 20$ , which is a typical number of key biomarkers that we hope will explain most of the observations. Since  $\delta \leq \sqrt{2} - 1$  is the compressibility condition for conventional sparsity, we take  $\delta = 1/4 = 0.25$ . We partition the set of 20,000 measurements into  $g = 6,000$  sets representing the number of pathways that we wish to study, and we take  $l_{\min} = 4$ , meaning that the shortest pathway of interest has four genes. Therefore we can take  $s_{\max} = \lfloor k/l_{\min} \rfloor = 5$ . Finally, let us take  $\zeta = 10^{-8}$ . With these numbers, it is readily verified that

$$m_S = 53,585, m_{GS} = 29,978.$$

In other words, both values of  $m$  are *larger than  $n$* ! Therefore one can only

conclude that these bounds for  $m$  are too coarse to be of practical use at least in computational biology, though perhaps they might be of use in other applications where  $n$  is a few orders of magnitude larger. Interestingly, the “deterministic” approach to the construction of  $A$  presented in [19] leads to smaller values of  $m$ , though in theory  $m$  increases as a fractional power of  $n$  as opposed to  $\log n$ . However, the method in [19] does not offer any advantage for group sparsity over conventional sparsity.

## 7 Conclusions

In this paper we have presented a unified approach for deriving upper bounds between the true but unknown sparse (or nearly sparse) vector and its approximation, when the vector is recovered by minimizing a norm as the objective function. The unified approach presented here contains the previously known results for  $\ell_1$ -norm minimization as a special case, and is also sufficiently general to encompass most of the norms that are currently proposed in the literature, including group LASSO norm, sparse group LASSO norm, and the group LASSO norm with tree-structured overlapping groups. Estimates for the number of measurements required are derived for group sparse vectors, and are shown to be smaller than for conventionally sparse vectors, when the measurement matrix is constructed using a probabilistic approach.

## Acknowledgement

The authors thank Mr. Shashank Ranjan for his careful reading of an earlier version of the papers.

## References

- [1] E. Candès, The restricted isometry property and its implications for compressed sensing, *Comptes rendus de l’Académie des Sciences, Série I* 346 (2008) 589–592.
- [2] M. A. Davenport, M. F. Duarte, Y. C. Eldar, G. Kutyniok, Introduction to compressed sensing, in: Y. C. Eldar, G. Kutyniok (Eds.), *Compressed Sensing: Theory and Applications*, Cambridge University Press, 2012, pp. 1–68.

- [3] S. Foucart, H. Rauger, *A Mathematical Introduction to Compressive Sensing*, Birkhäuser, 2013.
- [4] S. Negabhan, P. Ravikumar, M. J. Wainwright, B. Yu, A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers, *Statistical Science* 27(4) (December 2012) 538–557.
- [5] E. J. Candès, T. Tao, Decoding by linear programming, *IEEE Transactions on Information Theory* 51 (2005) 4203–4215.
- [6] D. Donoho, For most large underdetermined systems of linear equations, the minimal  $\ell_1$ -norm solution is also the sparsest solution, *Communications in Pure and Applied Mathematics* 59(6) (2006) 797–829.
- [7] A. Cohen, Wolfgang, Dahmen, R. DeVore, Compressed sensing and best  $k$ -term approximation, *Journal of the American Mathematical Society* 22(1) (January 2009) 211–231.
- [8] R. Tibshirani, Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society* 58(1).
- [9] M. Yuan, Y. Lin, Model selection and estimation in regression with grouped variables, *Journal of the Royal Statistical Society, Series B* 68 (2006) 49–67.
- [10] J. Huang, T. Zhang, The benefit of group sparsity, *The Annals of Statistics* 38(4) (2010) 1978–2004.
- [11] J. Friedman, T. Hastie, R. Tibshirani, A note on the group lasso and sparse group lasso, <http://www-stat.stanford.edu/tibs/ftp/sparsegrlasso.pdf> (2010).
- [12] N. Simon, J. Friedman, T. Hastie, R. Tibshirani, A sparse group lasso, <http://www-stat.stanford.edu/nsimon/SGLpaper.pdf> (2012).
- [13] R. Jenetton, J. Mairal, G. Obozinski, F. Bach, Proximal methods for hierarchical sparse coding, *Journal of Machine Learning Research* 12 (2011) 2297–2334.
- [14] G. Obozinski, L. Jacob, J.-P. Vert, Group lasso with overlaps: The latest group lasso approach, *arxiv* (2011) 1110.0413.
- [15] M. Bogdan, E. van den Berg, W. Su, E. J. Candès, Statistical estimation and testing via the sorted  $\ell_1$ -norm, <http://statweb.stanford.edu/candes/papers/SortedL1.pdf> (2013).

- [16] I. Daubechies, R. Devore, M. Fornasier, C. S. Güntürk, Iteratively reweighted least squares minimization for sparse recovery, *Communications on Pure and Applied Mathematics* 63(1) (2010) 1–38.
- [17] R. Baraniuk, M. Davenport, R. Devore, M. Wakin, A simple proof of the restricted isometry property for random matrices, *Constructive Approximation* 28 (2008) 253–263.
- [18] M. Vidyasagar, *Learning and Generalization: With Applications to Neural Networks and Control Systems*, Springer-Verlag, London, 2003.
- [19] R. DeVore, Deterministic construction of compressed sensing matrices, *Journal of Complexity* 23 (2007) 918–925.